POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

Executive Summary of the Thesis

Explainable Machine Learning and Deep Learning models to predict immunotherapy response in NSCLC patients using CT scans

Laurea Magistrale in Biomedical Engineering - Ingegneria Biomedica

Author: Margherita Favali

Advisor: Prof. Alessandra Laura Giulia Pedrocchi

Co-advisor: Vanja Miskovic, Arsela Prelaj, Alessandro Quarta

Academic year: 2022-2023

1. Introduction

According to estimates from the World Health Organization (WHO) cancer is the second leading cause of death globally. Lung cancer is classified in mainly two histological subtypes with different clinical behaviour: non-small cell lung carcinoma (NSCLC), accounting for 80-85% of all lung cancer cases, and small-cell lung carcinoma (SCLC). Even if traditional therapies like chemotherapy and radiotherapy provided benefit in terms of survival for lung cancer patients and are still incorporated in therapeutic algorithms, the prognosis remained poor with an estimated median overall survival (OS) of about 14 months in the metastatic setting. In this context, immunotherapy (IO) has brought a significant revolution in the treatment of NSCLC. In fact, recent clinical studies demonstrated that IO, delivered either alone or in combination with other therapies, could improve survival outcomes of advanced NSCLC patients, with about 20% of patients still alive 5 years after diagnosis of metastatic disease [1]. Hence, reliable biomarkers are required to identify patients that are most or least likely to benefit from this therapy. Since available biomarkers, such as PD-L1, demonstrated limited predicted efficacy, there is an urgent need for novel models to improve predictive capabilities. Analyzing CT scans using machine learning (ML) and deep learning (DL) techniques, offers a promising approach to extract features from medical images and construct predictive models. The present study develops a binary classification problem to characterize each patient if has or not a clinical benefit from IO. Furthermore, it assesses the predictive power of features coming from CT scans. The population involved in this retrospective study consisted of 375 patients from Fondazione IRCCS Istituto Nazionale dei Tumori with advanced NSCLC collected between April 2013 and May 2022. These patients received any-line anti-PD(L)1 therapy either alone or in combination with chemotherapy. Specifically, 305 patients were treated with IO, while 70 with the combination of IO and chemotherapy. Two different pipelines are implemented: the first is ML-based, while the second employs an end-to-end DL pipeline. The evaluation is performed on two data modalities: real-world data (RWD) and features extracted from CT scans. Significant effort is dedicated to ensure the explainability of both ML and DL models by using SHapley Additive exPlanations (SHAP).

2. Materials and methods

2.1. Data collection and curation

Two types of data were utilized: RWD and features extracted from CT scans. The RWD were collected during routine clinical exams at the baseline of IO. Clinical data selected for this study, based on clinicians hypothesis-driven, are 16 and they are collected in Table 1 with the relative description.

RWD	Definition
Therapy	Therapy administered to the patient: IO alone (1) or in combination with chemotherapy (0)
Age	Age of the patient at IO baseline
Sex	Patient sex: Male (1), Female (0)
Surgery y/n	Binary variable identifying patients which underwent surgery to reduce tumor mass
Histology	Binary variable for tumor type: squamous (1) or non squamous (0)
Line of therapy	Line of treatment that patient received
Smoking status	Binary variable that identifies if the patient is a smoker or ex-smoker (1) or non-smoker (0)
PDL1 group	Value of Programmed death ligand 1 (PD-L1): < 1% (1), 1-49% (2), >50% (3)
ECOG PS	ECOG Performance Status at IO baseline
Tumor stage	Tumor characterization according to TNM evaluation
Node stage	Node characterization according to TNM evaluation
Metastases stage	Metastasis characterization according to TNM evaluation
N of metastatic sites	Indicates the number of the metastases
Metastases Brain	Binary variable that indicates brain metastasis
Metastases Bone	Binary variable 7that indicates bone metastasis
BMI	Body Mass Index at IO baseline

Table 1: RWD

An extensive data curation procedure was performed. Duplicate patients and inconsistent values were eliminated and missing data were filled with the assistance of clinicians, whenever possible. Textual data were converted into numerical and categorical values, and imputation techniques were employed to address any remaining missing data. The other data used were features extracted from the primary tumor volume of patients' baseline CT scans. In ML pipeline, they were calculated with pyradiomics package and encompassed shape characteristics, grey level properties, grey tone differences, and statistical attributes [2]. In DL approach, features were directly extracted from the neural network.

2.2. Outcome

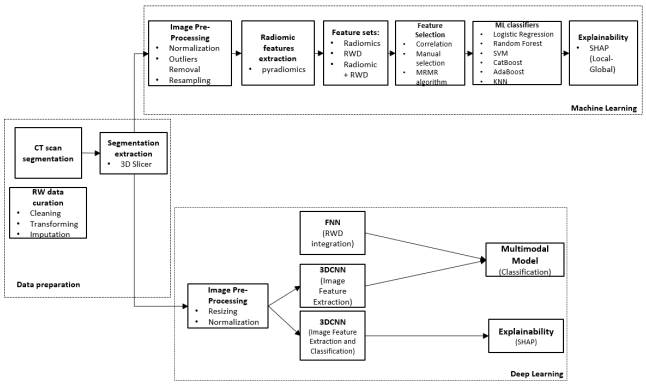
The target value is represented by the best overall response, i.e. the best response recorded from the first radiological evaluation until disease progression according to the Response Evaluation Criteria in Solid Tumors (RECIST1.1) [3]. The outcome analyzed is the Clinical Benefit Rate (CBR), which is defined as the percentage of patients who have achieved complete response (CR), partial response (PR), or at least four months of stable disease (SD) as a result of therapy. The outcome takes into account also patients that had a progression but with still a clinical benefit after at least 9 months. Two classes were defined as follows:

- Class 0: Progressive Disease (PD) if TTF $<$ 9 months, Stable Disease (SD) if PFS $<$ 4 months
- Class 1: Complete Response (CR), Partial Response (PR), Stable Disease (SD) if PFS $\geq$ 4 months, Progressive Disease (PD) if TTF $\geq$ 9 months

The choice of using a clinical endpoint rather than the radiologic alone was done after discussions with clinicians, as this endpoint can better distinguish patients who benefit from IO from refractory patients.

2.3. Model development

Two different approaches were used: one ML-based and the other DL-based, as Figure 1 shows.



(a)

Data modality	Machine Learning	Deep Learning
CT scan features	Yes	Yes
RWD	Yes	No
CT scan features + RWD	Yes	Yes

(b)

Figure 1: Two approaches: (a) Machine Learning and Deep Learning Illustration of methodological workflow

2.3.1 ML approach

The first approach is a classification using ML pipeline (Fig. 1). It started with the segmentation extraction from the CT scans by using 3D slicer, an open source software for visualization, processing and segmentation of medical images. Subsequently, several steps were undertaken to pre-process the images and extract the radiomic features from the CT scans [2]. 107 radiomic features were extracted, which were categorized into seven different classes: 18 features of the firstorder class, 14 shape descriptors, 75 texture features. Then a three-step feature selection process was implemented to eliminate redundant information in the dataset. The steps involved: (1) correlation analysis, (2) manual selection, and (3) Maximum Relevance- Minimum Redundancy (MRMR) technique. MRMR algorithm aims at selecting the features that have maximum relevance with respect to the target variable and minimum redundancy with respect to the features selected at previous iterations [4]. Six ML classifiers (Logistic Regression, Random Forest, SVM, CatBoost, AdaBoost, KNN) were fed with three different feature sets: radiomic features, combination of radiomics and RWD, and RWD. The best-performing classifier was identified for each feature set and it was tested on an external validation set. SHAP was employed to explain model predictions on the test set, identifying the features that had the greatest impact on the outcome and understanding how they influenced it [5]. Two approaches were used to provide the explainability: global, to understand how a model made all the predictions, and local, for understanding how the model made decisions for a single prediction. For the global solution, SHAP summary plot was exploited. For the local solution, waterfall plots were generated for four types of predictions: True Positive, True Negative, False Positive, and False Negative. Features are ordered from top to bottom based on their importance, and the contribution of each feature to the individual prediction is displayed. Features that move the prediction towards class 1 are represented by red bars, while features that contribute to predict class 0 are the blue bars.

2.3.2 DL approach

The second approach included the implementation of two end-to-end solutions (Figure 1) of DL pipeline. End-to-end means that the model learns to automatically extract relevant features and make classification in a single integrated process. The first pipeline is a 3D Convolutional Neural Network (3DCNN) that solely processes DL features coming from the images. The second is a bimodal model that receives a combination of both DL features and RWD as input. After that, the neural models were trained and tested with the same training, test and external validation sets used for ML. In case of model with images only, local SHAP values were used to explain the model predictions. Since features are essentially pixels in DL, model explainability helps to identify pixels which contribute negatively or positively to the predicted class. Important pixels for the prediction are assigned colors: red pixels represent positive SHAP values that contributed to classify image in class 1, while blue pixels represent negative SHAP values that contributed to classify image in class 0.

3. Results

3.1. Dataset

The present study involves a cohort of 375 patients with NSCLC treated with IO as any-line of therapy for advanced disease. 236 patients were used as the training set, while 59 patients were allocated to the test set. Additionally, 80 patients were reserved as an external validation set exclusively for the best-performing model.

3.2. Classification with ML

3.2.1 Radiomic features set

Once radiomic features were extracted from CT scans, the best set of features to feed the ML models was selected through a three-step feature selection. The first step involved checking for highly correlated features in order to remove them as they carry nearly identical information. Out of the initial 107 radiomic features, 77 were found to be highly correlated, leaving 30 features. Subsequently, remaining features that did not differentiate between class 1 and class 0, were eliminated, resulting in a set of 21 radiomic

features. In the last step, MRMR was implemented to choose the optimal number of features to feed each ML classifier. The best-performing model resulted the Logistic Regression (LR). It was fed with 15 radiomic features and it achieved accuracy = 0.61 and AUC = 0.58. To assess the robustness and performance of this model, an external validation set was utilized and the resulting accuracy and AUC were 0.54. Figure 2 shows how the features impact globally the predictions on the test set. Features are arranged on the y-axis based on their importance for the model outcome, with the most important feature positioned at the top. On the x-axis, the plot indicates whether the effect of the feature value is associated with class 1 or class 0. A color map is used to represent the feature values, where red indicates high values and blue low values.The feature that has the highest influence in the model's outcome is Major Axis Length.

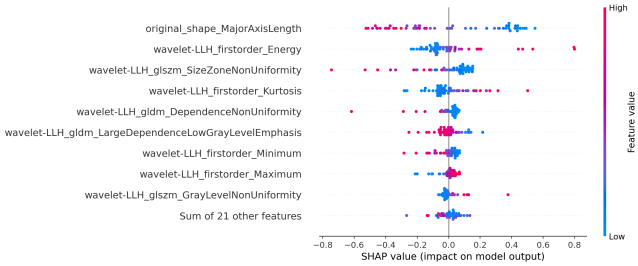


Figure 2: Radiomics: Global Explainability of LR on test set

It belongs to the shape class, where features are descriptors of the three-dimensional size and shape of the ROI and are independent from the gray level intensities distribution. This feature measures the largest axis length of the ROI-enclosing ellipsoid. Specifically, lower values of this feature move the prediction towards class 1, while higher values are more likely to be present in class 0.

3.2.2 RWD

The correlation matrix including RWD was plotted and it demonstrated that none of the RWD were correlated with each other. Therefore, all of them were utilized as input for the MRMR algorithm, without any additional manual selection before. With MRMR implementation, optimal number of features for each model was found. SVM classifier fed with 15 features, demonstrated the most promising results on RWD dataset with accuracy = 0.68 and AUC = 0.71. Global explainability with SHAP on the test set is shown in Figure 3.

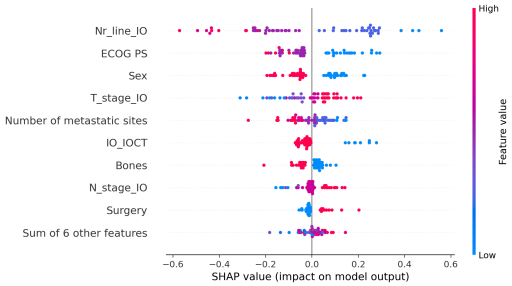


Figure 3: RWD: Global Explainability of SVM on test set

The feature that has highest influence in the model's outcome is Line of therapy. Specifically, lower values of this feature move the prediction towards class 1, while higher values move the prediction towards class 0.

3.2.3 Combination of radiomics and RWD

The same set of 21 radiomic features, obtained after applying the procedure described in Section 3.2.1 to remove highly correlated and poor informative features, and the 16 baseline RWD were included in this dataset, resulting in 37 features. The MRMR feature selector was then applied to select the optimal number of features for each model. LR performed better, achieving accuracy = 0.69 and AUC = 0.73 with 25 features. Subsequently, on the external validation set, performance metrics were: accuracy = 0.69 and AUC = 0.71. The SHAP values for LR were computed. Figure 4 shows the global explanation on the test set. Global SHAP revealed that the two features that most influenced the predictions were ECOG PS (RWD) and Large Dependence Emphasis (LDE), a radiomic feature. For ECOG PS, higher values shifted the predictions towards class 1, which is a confirmation of what is assessed by clinical practice, as a lower value of the ECOG PS scale indicates better patient clinical condition. LDE can be translated into a measure of the texture of the lesion, where higher values indicate a more homogeneous texture. The Figure 4, reveals that high values for LDE are correlated with class 1, suggesting

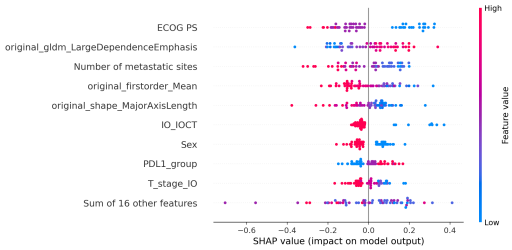


Figure 4: Radiomics and RWD: Global explainability of LR on test set

that a more homogeneous texture is most likely to correspond to a patient belonging to class 1. Figure 5 shows two examples of CT scans, where high value of LDE is clearly associated with more homogeneous texture.

3.3. Classification with DL

3.3.1 DL features

3DCNN consisted of three convolutional layers, each followed by a rectified linear unit, a max pooling layer, and a dropout layer. Additionally, there were two linear layers, with the final layer dedicated to binary classification. Loss and accuracy for both training and test sets were computed. Based on both metrics, overfitting was present, meaning that the model performed well during training but poorly on test data, which represents unseen data. In case of accuracy, the majority of epochs showed significantly lower test accuracy compared to the training accuracy. The train loss demonstrated a consistent pattern of decreasing as the number of epochs increases. However, the test loss exhibited an irregular behavior, increasing instead of decreasing from left to right. The best value for test accuracy was 0.63, reached at epoch 42, where the test loss was 0.887. In the external validation set, accuracy was 0.55 and loss was 0.913. Local explanation on the test set was carried out using SHAP.

3.3.2 DL features and RWD

In the bimodal model, two data modalities (RWD and DL features) were processed. Two neural networks were implemented: a 3DCNN, slightly different from the previous one, for processing the 3D images and extract relative features, and a Feed-Forward neural network (FNN) for handling the RWD. The features pertaining to CT scans and RWD were extracted from the two neural networks and concatenated within the bimodal model, where the classification process was carried out (intermediate fusion). Loss and accuracy for training and testing were computed. For the test set, the highest accuracy reached was 0.64, achieved at epoch 60, where loss was 0.170. In the external validation, accuracy resulted 0.65 and loss was 0.119.

4. Discussion

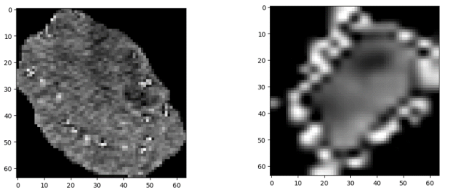
4.1. ML feature sets comparison

One of the objectives of this study was to evaluate the predictive capability of radiomic features, either alone or in combination with RWD. Three different feature sets were used to fed six ML classifiers. The best performing model was selected for each feature set: LR was the best for both the radiomics and combination, while SVM was chosen for RWD. Radiomics alone were not efficient in the classification (acc = 0.61), while RWD demonstrated greater robustness and efficiency in predicting therapy response (acc = 0.68). However, the results of this study revealed that the addition of radiomics to RWD did not add significantly more information compared to a prediction model based solely on RWD. In fact, accuracy reached with the combination was 0.69, while RWD got 0.68. AUC metric confirmed the low predictive power of radiomics alone (AUC = 0.58), while a similar performance in combination (AUC = 0.73) and RWD (AUC = 0.71).

4.2. Explainability analysis results

To provide an explanation of how both radiomics and RWD were utilized to predict the response, SHAP Summary Plot of LR trained on radiomics and RWD (Fig. 4) is considered. It is important to note that the model achieved an accuracy of 0.69 and an AUC of 0.73. These results indicate that further work is required to improve the model's performance and reliability. Although the AUC of 0.73 is considered acceptable, it does not reach the level of excellence, particularly for medical applications, where AUC should be higher than 0.8. Given that performance is not outstanding, the reliability of its explainability may be compromised. Consequently, the interpretability of the model may be biased towards certain features and their corresponding value explanations. ECOG PS

(RWD) emerged as the most influential feature. Additionally, the treatment received by the patients also provided valuable information. Patients who received a combination of IO and chemotherapy have a higher probability of positive response. This is confirmed by many studies in clinical knowledge. SHAP Summary Plot revealed interesting insights into the radiomic features, which can be confirmed by reviewing the CT scans. An association between LDE in terms of radiomic meaning and texture homogeneity was observed. In particular, CT scans with high LDE (Figure 5a) show homogeneous textures, while CT scans with low LDE (Figure 5b) show a non homogeneous texture (pixel variations).



(a) LDE = 581.914 (b) LDE = 257.381

Figure 5: CT scans: (a) High LDE (b) Low LDE

In the DL pipeline, SHAP values provided initial insights into the functioning of neural networks. It seems that the network primarily focuses on the edges of the ROI, while the internal region of the ROI does not significantly influence the prediction. However, further analysis is necessary to fully comprehend which tumor characteristics contribute to specific predictions.

4.3. ML and DL comparison

One of the objectives of this study was to evaluate the performance of ML and DL techniques (Table 2) and determine which approach could be superior in predicting IO response in this clinical context. Since the dataset was balanced with respect to classes in the test set, the performances can be compared based on test accuracy results.

When considering models that uses features derived from CT scans, DL (acc = 0.63) on the test set demonstrates slightly higher efficiency compared to ML (acc = 0.61). However, when incorporating both RWD and features derived from CT scans, ML technique (acc = 0.69) outperforms DL approach (acc = 0.64). Furthermore, when combining CT scans features with

Data modality	Machine Learning Accuracy		Deep Learning Accuracy	
	Test	Validation	Test	Validation
CT scan features	0.61	0.54	0.63	0.55
RWD	0.68	0.7	No	
CT scan features + RWD	0.69	0.69	0.64	0.65

Table 2: Machine Learning and Deep Learning performance comparison

RWD, ML techniques exhibit an increase in their predictive performance. Indeed acc = 0.61 is achieved with radiomics only, while acc = 0.69 with the combination. However, this does not happen for DL, where performance achieved using CT scan features (acc = 0.63) and performance with the combination (acc = 0.64) can be considered comparable. Considering external validation, more significant improvement when adding RWD to CT scan features is observed.

4.4. Limitations and Future research

Finally, none of the ML and DL approaches with the present data types yielded satisfactory results that would allow the model to be applied in possible clinical practice. There could be several reasons for this. Firstly the population included an heterogeneous cohort of patients treated with IO in a different treatment lines and a wide range of time (2013–2023), when different CT image acquisition protocol were applied. In addition, CT scan exams were performed at different Institutions. These two considerations could have produced some intrinsic noise during the feature extraction. Furthermore, the limited number of patients included in the study could have influenced the poor results. In both ML and DL pipelines, including a bigger and more homogeneous cohort of patients to the current dataset would improve performance. Secondly, the number of features utilized could be expanded. The current image pre-processing and feature extraction and selection methods may not be the optimal solutions for this type of problem. In the ML approach, it is important to note that no specific filter was applied to extract the features. However, a broader range of features could be extracted by employing various types of filters that have the potential to capture different aspects of the

underlying data, and homogenized dataset coming from different institutions. This approach would enable the inclusion of a larger quantity of radiomic features that could be explored in terms of their predictive capabilities. For both ML and DL approaches, the implementation of additional image preprocessing techniques can enhance the quality of input images and contribute to more accurate feature extraction. The third main limitation regards the request to completely understanding clinical problem. It is not sufficient to apply ML and DL solely for predicting response outcomes. Survival outcomes, such as progression-free survival (PFS) and overall survival (OS), should be used since they are more relevant clinical outcomes. Additionally, the combination of RWD and CT scans with other data types, such as genomics and digital pathology, could provide better insights into the clinical problem. This is supported by numerous studies in the literature, which demonstrated the improvements achieved by integrating features from different data modalities (histopathological, radiomic, genomic, and clinical data) into multimodal models. Another potential solution, considering the data types used in the present study, is the utilization of delta-radiomic features. These features represent the differences between radiomics extracted from two CT scans: the baseline CT scan and the post-baseline CT scan. Incorporating delta-radiomic features may provide valuable information on the changes in radiomic characteristics over the course of treatment. Last main problem concerns DL approach, where preliminary results were found; to improve performance, transfer learning techniques and the exploration of different neural network architectures could be employed.

5. Conclusions

The objective of this study was to identify radiomic and clinical biomarkers associated with IO benefit in a cohort of patients with NSCLC. Medical applications require very high reliability and performances in order to be applied in clinical practice. These initial achievements could be the base of the ambitious but ultimate goal of developing novel tools for the selection of ideal candidates for IO. So by investigating and incorporating future perspectives, this research may have the potential to contribute to the development of innovative approaches that can be applied in clinical practice.

References

[1] Marina C Garassino, Shirish Gadgeel, Giovanna Speranza, Enriqueta Felip, Emilio Esteban, Manuel Dómine, Maximilian J Hochmair, Steven F Powell, Helge G Bischoff, Nir Peled, et al. Pembrolizumab plus pemetrexed and platinum in nonsquamous non–small-cell lung cancer: 5-year outcomes from the phase 3 keynote-189 study. *Journal of Clinical Oncology*, 41(11):1992, 2023.

[2] Pyradiomics. `https://pyradiomics.readthedocs.io/en/latest/`, 2016.

[3] Elizabeth A Eisenhauer, Patrick Therasse, Jan Bogaerts, Lawrence H Schwartz, Danielle Sargent, Robert Ford, Janet Dancey, S Arbuck, Steve Gwyther, Margaret Mooney, et al. New response evaluation criteria in solid tumours: revised recist guideline (version 1.1). *European journal of cancer*, 45(2):228–247, 2009.

[4] Milos Radovic, Mohamed Ghalwash, Nenad Filipovic, and Zoran Obradovic. Minimum redundancy maximum relevance feature selection approach for temporal gene expression data. *BMC bioinformatics*, 18(1):1–14, 2017.

[5] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.